Information-theoretic bound on the energy cost of stochastic simulation

Karoline Wiesner,[1,*] Mile Gu,[2] Elisabeth Rieper,[2] and Vlatko Vedral[3,4,2]

[1] *School of Mathematics, Centre for Complexity Sciences, University of Bristol, University Walk, Bristol BS8 1TW, United Kingdom*
[2] *Centre for Quantum Technologies, National University of Singapore, 3 Science Drive 2, S15-03-18, Singapore 117543, Singapore*
[3] *Atomic and Laser Physics, Clarendon Laboratory, University of Oxford, Parks Road, Oxford OX1 3PU, United Kingdom*
[4] *Department of Physics, National University of Singapore, 2 Science Drive 3, Singapore 117543, Singapore*

(Dated: October 20, 2011)

Physical systems are often simulated using a stochastic computation where different final states result from identical initial states. Here, we derive the minimum energy cost of simulating a complex data set of a general physical system with a stochastic computation. We show that the cost is proportional to the difference between two information-theoretic measures of complexity of the data – the *statistical complexity* and the *predictive information*. We derive the difference as the amount of information erased during the computation. Finally, we illustrate the physics of information by implementing the stochastic computation as a Gedankenexperiment of a Szilard-type engine. The results create a new link between thermodynamics, information theory, and complexity.

arxiv.org/phys/0905.2918

PACS numbers: 02.50.Ey Stochastic processes, 05.70.-a Entropy thermodynamics, 89.70.Cf Entropy information theory, 89.70.-a Information theory, 89.75.-k Complex systems,

The idea of physics as information has a long history. The concept of entropy, at the heart of information theory, originated in the theory of thermodynamics. It was Maxwell and Boltzmann who, in the beginning of the 19^{th} century, recognized the intricate link between probability distributions over configurations and thermodynamics. This laid the foundation to the field of statistical mechanics. The similarity between the thermodynamic entropy and the information entropy, introduced in 1948 by Shannon, lead to a whole new perspective on physical processes as storing and processing information. It also lead to paradoxes such as Maxwell's demon which seemed to suggest that work could be generated from heat only with the use of information, which would violate the second law of thermodynamics (for a review, see Refs. [1, 2]). The paradox was solved independently by Penrose and Bennett, in considering the entropy creation caused by erasing information [1, 3].

With the insight that erasing information generates entropy, Zurek found limits on the thermodynamic cost of deterministic computation using algorithmic complexity [4]. A *deterministic computation* generates a unique output given a particular input. Repeated computations yield identical results. A *stochastic computation* on the other hand yields different outputs for identical inputs. It is a useful descriptor of natural processes which are often stochastic and have different final states given 'identical' initial states (within a given finite resolution or, in quantum mechanics, even infinite resolution). Here, we derive the minimum energy cost of simulating a complex data set with a stochastic computation. We show that it is proportional to the difference between the *statistical complexity* [5] and the *predictive information* of the data [6]. We derive this difference as the amount of information erased during the computation. Finally, we illustrate the physics of information by "implementing" the stochastic computation with a Szilard-type engine.

It has been shown that the difference between these two measures arises from an asymmetry in the transport of information forward and backward in time [7]. In this paper we give a physical explanation of this asymmetry together with new mathematical proofs of the relevant information theory.

Consider the following model of stochastic computation: A given computational device is in some initial state and outputs length-N strings of symbols x^N according to probability distribution $\Pr(X^N)$. If the distribution $\Pr(X^N)$ is statistically indistinguishable from that of an observed data set of a physical system the symbol sequence x^N is a simulation of that system. Where the probability distribution $\Pr(X^N)$ is a very uncompressed description, one step away from raw data of an experiment, the computational device simulating it is a very compact description, a summary of the regularities, a first step toward a 'theory' explaining an experiment. We call the joint probability distribution of past and future observations $\Pr(X_{-N}^{N-1})$ a *stochastic process*. The provably unique minimal (in terms of information stored) and optimal (in terms of prediction) such computation-theoretic representation summarising the regularities of a stochastic process is a so-called ϵ-machine [5, 8]. It consists of an output alphabet $\mathcal{A}$, a set of *causal states* $\mathcal{S}$ and stochastic transitions between them. For every pair of states $s, s' \in \mathcal{S}$ probabilities $\Pr(S_i = s' | S_{i-1} = s, X_i = x)$

are defined for going from state s to state s' while outputting symbol $x \in \mathcal{A}$. The *statistical complexity* of a process is defined as the Shannon entropy over the stationary distribution of its ϵ-machine's causal-states [22]:

$$C_\mu := H(\mathcal{S}) \ . \tag{1}$$

C_μ is the number of bits required to specify a particular causal state in the ϵ-machine. It is the number of bits that need to be stored to optimally predict future data points.

The *predictive information* of a data set is given by the mutual information between the two halves, e.g. the past data and the future data [8, 9]:

$$\mathbf{E} = \lim_{N\to\infty} I[X_{-N}^{-1}; X_0^{N-1}] \ . \tag{2}$$

where X_{-N}^{-1} and X_0^{N-1} are strings of random variables representing observations of a stochastic process. Predictive information is also known under the name of excess entropy, effective measure complexity, and stored information (see [9] and refs. therein). For the following thermodynamic development of stochastic computation we find the name predictive information most suitable. The predictive information, measured in bits, can be interpreted as the average number of bits a process stores at a given point in time and "transmits" to the future. It is known that

$$C_\mu = \mathbf{E} + H(S_{-1}|X_0^\infty) \tag{3}$$

and hence that $\mathbf{E} \leq C_\mu$ [8, Theorem 5]. Two important properties of ϵ-machines of relevance here are that the next state given the last state and the current symbol is uniquely determined (Eq. 4) and that the state after the observation of a long enough sequence of symbols is uniquely determined (Eq. 5) [8]:

$$H(S_N|S_{N-1}X_N) = 0 \quad \text{(deterministic-stochastic)} \tag{4}$$

$$\lim_{N\to\infty} H(S_N|X_0^N) = 0 \quad \text{(synchronising)} \tag{5}$$

Successful inference of ϵ-machines ranges from dynamical systems [5], spin systems [10], and crystallographic data [11] to molecular dynamics [12], atmospheric turbulence [13], and self-organisation [14].

Landauer defines an operation to be *logically irreversible* if the output of the operation does not uniquely define the inputs [15]. In other words, logically irreversible operations erase information about the computational device's preceding logical state. Landauer's insight was that logical information erasure costs energy [15]. In the following we discuss how Landauer's principle and logical irreversibility apply to stochastic computation and, in particular, to the computation of a stochastic process. For a given ϵ-machine the current state and the next symbol determine the next state uniquely (Eq. 4). The reverse, however, is not necessarily true. Given the current state and last output symbol, the previous state is not always uniquely determined. In this case the ϵ-machine is *logically irreversible*. Following Landauer's definition of irreversibility, we define the information erasure per computational step of a given ϵ-machine as the entropy of the previous state (S_{i-1}) given the current state (S_i) and last output symbol (X_i) :

$$h_{erase} := H(S_{i-1}|X_i S_i), \tag{6}$$

For later use, we also define $h_{erase}^N := H(S_{i-N}|X_{i-N}^i S_i)$ for $i > N$ and $H_{erase} := \lim_{N\to\infty} h_{erase}^N$. h_{erase} can be calculated from the ϵ-machine directly. It quantifies the average *irreversibility* of a computational step of the ϵ-machine. We now show that strict equality between $\mathbf{E}$ and C_μ holds if and only if the ϵ-machine is fully logically reversible, i.e. *iff* $h_{erase} = 0$.

Theorem 1. *The predictive information* $\mathbf{E}$ *of a stochastic process is equal to the statistical complexity* C_μ *if and only if the information erasure of the corresponding* ϵ*-machine is zero:*

$$C_\mu = \mathbf{E} \Leftrightarrow h_{erase} = 0 \ . \tag{7}$$

Proof. **'⇒':**

From the Markov property of the states S_i (i.e. $H(X_1|S_{-1}X_0S_0) = H(X_1|X_0S_0)$) it follows that $H(S_{-1}|X_0X_1S_0) = H(S_{-1}|X_0S_0)$. Using this recursively and the fact that further conditioning never increases the entropy we obtain the forward direction

$$\begin{aligned} C_\mu = \mathbf{E} &\Leftrightarrow H(S_{-1}|X_0^\infty) = 0 \\ &\Rightarrow H(S_{-1}|X_0^\infty S_0) = 0 \\ &\Rightarrow H(S_{-1}|X_0 S_0) = 0 \end{aligned} \tag{8}$$

'⇐': Going in reverse we have $H(S_{N-1}|X_N S_N) = 0 \ \Rightarrow H(S_{N-1}|X_0^N S_N) = 0$ and in addition

$$H(S_{N-1}|X_0^N S_N) = H(S_{N-1}|X_0^N) - H(S_N|X_0^N) \ . \tag{9}$$

The second term on the RHS goes to zero in the limit $N \to \infty$ (Eq. 5). In the same way $H(S_{N-k}|X_0^N S_{N-k+1}) \to H(S_{N-k+1}|X_0^N)$, $k = 2, 3, \ldots, N+1$, as $N \to \infty$. Setting $k = N+1$, the claim follows. □

Note that this result automatically implies that any ϵ-machine with $h_{erase} = 0$ can be inverted in time, turning it into a *retrodicter* as introduced in [7] and thus providing an immediate and quick construction of a retrodicter for this case saving the 'modeler' a second computationally costly inference procedure (for more on computational cost see e.g. [16]).

For future reference we call $H(X_i|S_i) := h^R$ the uncertainty of the last symbol given the current state. This reverse entropy rate is different from the reverse

entropy rate referred to in dynamical systems theory (see e.g. [17]). The complimentary quantity to h^R is the well-known entropy rate of a stochastic process $h = H(X_{i+1}|S_i)$. We can see that the amount of information which can be erased per computational step is upper bounded by the amount of information which is created per computational step,

$$h_{erase} \leq h \ , \tag{10}$$

by writing the joint entropy $H(S_{i-1}X_iS_i)$ as two different sums:

$$\begin{aligned} H(S_i|S_{i-1}X_i) + H(S_{i-1}X_i) &= H(S_{i-1}|X_iS_i) + H(X_iS_i) \\ \Leftrightarrow \quad & H(X_i|S_{i-1}) + H(S_{i-1}) \\ & = H(S_{i-1}|X_iS_i) + H(X_i|S_i) + H(S_i) \\ \Leftrightarrow \quad & h^R = h_{erase} + h \ , \end{aligned} \tag{11}$$

where, in the second line, we have used Eq. 4.

Now consider a ϵ-machine $\mathcal{M}$ with $h_{erase} > 0$. To derive the difference between $\mathbf{E}$ and C_μ we construct an ϵ-machine which outputs more than one symbol at a time as follows. For every pair of states $s_i, s_j \in \mathcal{S}$ of ϵ-machine $\mathcal{M}$ we construct the N^{th} concatenation $\mathcal{M}^{\otimes N}$ with state transition probabilities $\Pr(s_j|s_ix^N) = \sum_{s_1^{N-1} \in S^{N-1}} \Pr(s_j|s_{N-1}x_N) \left[\prod_{k=2}^{N} \Pr(s_k|s_{k-1}x_k)\right] \cdot \Pr(s_1|s_ix_1)$ upon outputting x^N, where the sum runs over all state sequences s_1^{N-1} of length $N-1$. Note, that $\mathcal{M}$ and $\mathcal{M}^{\otimes N}$ have the same set of states and the same probability distribution over output strings $\Pr(X^N)$ – so they have the same $\mathbf{E}$ and C_μ.

Theorem 2. *The difference between the statistical complexity C_μ and the predictive information $\mathbf{E}$ of a stochastic process approaches the information erased by the corresponding concatenated ϵ-machine $\mathcal{M}^{\otimes N}$ as $N \to \infty$:*

$$\lim_{N\to\infty} h_{erase}^N = C_\mu - \mathbf{E} \ .$$

Proof. By rewriting $H(S_{-1}X_0^NS_N)$ in two different ways we obtain the information-theoretic equality:

$$\begin{aligned} H(S_{-1}|X_0^N) =& H(S_{-1}|X_0^NS_N) \\ & + H(S_N|X_0^N) - H(S_N|S_{-1}X_0^N) \end{aligned} \tag{12}$$

The last term is zero due to determinism (Eq. 4), the second term goes to zero for $N \to \infty$. Hence, taking the limit we obtain:

$$H(S_{-1}|X_0^N) \to H(S_{-1}|X_0^NS_N) \text{ as } N \to \infty \ . \tag{13}$$

The term on the right-hand side is exactly h_{erase} of $\mathcal{M}^{\otimes N}$. Letting $N \to \infty$ the claim follows. □

Using Theorem 2 we can derive the minimum energy cost of simulating a stochastic process. Fig. 1 schematically illustrates an ϵ-machine contained in a box which on consecutive time steps outputs symbols visible to an outside observer. The computational steps are as follows. In (a) the ϵ-machine in the box is in state S_{i-1}. Going to (b) it generates symbol X_i according to the probability distribution $\Pr(X_i|S_{i-1})$, leading to an increase in entropy inside the box by $h = H(X_i|S_{i-1})$. Going to (c) the ϵ-machine moves from state S_{i-1} to state S_i. Erasure of the previous-state information causes a decrease in entropy inside the box by $H(S_{i-1}|S_iX_i) = h_{erase}$. Finally, in (c) the symbol is ejected into the environment which decreases the entropy inside the box again, this time by $H(X_i|S_i) = h^R$. This closes one cycle of computation. The entropy contributions during one closed cycle must add up to zero and we obtain $h - h_{erase} - h^R = 0$ which is exactly Eq. 11.

We now modify this stochastic computation by allowing for the generation of $N+1$ symbols at a time. The ϵ-machine inside the box (Fig. 1) is replaced by the concatenated ϵ-machine $\mathcal{M}^{\otimes(N+1)}$. In (a) this machine starts out in state S_{i-1}, going to (b) it generates $N+1$ symbols according to $\Pr(X_i^{i+N}|S_{i-1})$. This causes an increase in entropy inside the box by $H(X_i^{i+N}|S_{i-1})$. Going to (c) the concatenated ϵ-machine moves to state S_{i+N} erasing the previous-state information which decreases the entropy inside the box by $H(S_{i-1}|X_i^{i+N}S_{i+N})$. Ejecting the symbol sequence into the environment the entropy of the box decreases by $H(X_i^{i+N}|S_{i+N})$. Setting w.l.o.g. $i = 0$, we obtain for the entropy balance of one computational cycle:

$$H(X_0^N|S_{-1}) - H(S_{-1}|X_0^NS_N) - H(X_0^N|S_N) = 0 \ . \tag{14}$$

The LHS can be rewritten as

$$H(S_{-1}) - H(S_{-1}|X_0^NS_N) - I(S_{-1};X_0^N) \tag{15}$$

Letting $N \to \infty$ we obtain

$$H_{erase} + \mathbf{E} - C_\mu = 0 \ . \tag{16}$$

Hence, the entropy balance of one cycle of stochastic computation is in the limit of an infinite string of output given by Eq. 3 and Theorem 2. We now discuss the thermodynamics of such a stochastic computation.

The Carnot efficiency of one cycle of an engine consisting of an ideal gas in a cylinder alternately connected to a hot and a cold reservoir at temperature T_H and T_C, respectively, is, as is well known, given by the ratio of work output W and absorbed heat Q_H:

$$\eta = \frac{W}{Q_H} = 1 - \frac{T_C}{T_H} \ . \tag{17}$$

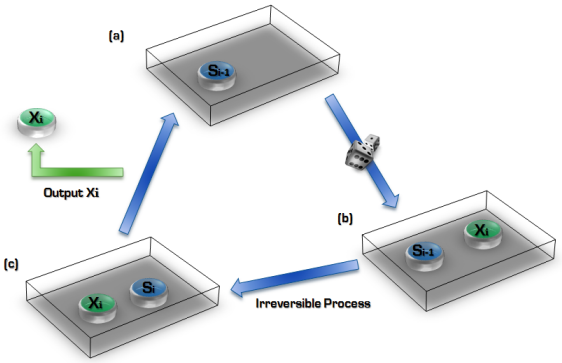

FIG. 1: One computational cycle for simulating a stochastic process with an ϵ-machine.

In 1929, Szilard invented a Gedankenexperiment of a single-particle engine to resolve Maxwell's demon paradox which seemed to defy the second law of thermodynamics [18]. The engine consists of a particle in a box, a measurement device locating the particle in either half of the box, and a memory to store the measurement result. Szilard considered the following procedure for extracting work from the particle's thermal motion. A user measures the particle's position and stores this one bit of information in the memory. Subsequently she 'compresses' the box to the half which contains the particle. This does not require any work. The thermal motion of the particle 'decompresses' the box again and hence lets the user extract work from the box without any cost. This apparent paradox is resolved when one factors the additional energy required to for the user to reset her memory [15]. With the memory initially at temperature T_H and the reset done at temperature T_C the efficiency of this engine is given by the Carnot efficiency (Eq. 17) and the laws of thermodynamics were restored.

This Gedankenexperiment can be extended to a particle in one of $2^{\mathbf{E}}$ possible partitions. Hence, storing the measurement result of the particle's position requires E bits of memory. Following the same argument as for the original Szilard engine we obtain Carnot efficiency $\eta = 1 - T_C/T_H$. This modified engine leads us directly to a new interpretation of Theorem 2.

In the simulation of a stochastic process, we attempt to generate information about its future based on observations of the past. This may be viewed as a Gedankenexperiment where we attempt to maximize our knowledge about a particle whose position is governed by a random variables $X_0, X_1, ...$ which we can only indirectly measure by recording appropriate information from a correlated random variables $X_{-1}, X_{-2}, ...$ These recorded bits can the be translated to information about $X_0, X_1, ...$ by the use of an appropriate simulator. To extract the maximum possible amount of information, $\mathbf{E}$, the theory of ϵ-machines dictates that we must record at least record C_μ bits. The minimality and optimality of ϵ-machines ensures that any fewer bits would render the simulation sub-optimal. This results in a stochastic computation that allows to extract

$Q_H^{sc} = kT_H \mathbf{E} \ln 2$ units of extra work. Meanwhile, the C_μ bits stored about the past are erased, which costs $Q_C^{sc} = kT_C C_\mu \ln 2$ units of energy. The efficiency is, just like before, the ratio between output work and absorbed heat:

$$\frac{W^{sc}}{Q_H^{sc}} = 1 - \frac{Q_C^{sc}}{Q_H^{sc}} = 1 - \frac{C_\mu T_C}{\mathbf{E} T_H} \ . \tag{18}$$

We define the information-theoretic efficiency for computing a stochastic process:

$$\iota := \frac{\mathbf{E}}{C_\mu} = 1 - \frac{H_{erase}}{C_\mu} \ . \tag{19}$$

Combining the thermodynamic and information theoretic efficiencies we obtain

$$\frac{W^{sc}}{Q_H^{sc}} = \eta(\iota) = 1 - \iota^{-1} \frac{T_H}{T_C} \ . \tag{20}$$

For maximal information-theoretic efficiency we recover the thermodynamic efficiency from Eq. 17.

$\mathbf{E}/C_\mu$ has been named the "predictive efficiency of a process as the fraction of the information it contains which actually effects the future" [19]. Our results supply this concept with physicality and mathematical rigour.

We have derived the minimum energy cost of simulating a physical system as the difference between two information-theoretic complexity measures of the data. Of the two measures, the predictive information measures the amount of information stored about a process's past transmitted to the future, the statistical complexity measures the amount of information required to compute this future. Any difference between the two is given by the amount of information erased during the simulation of the data and hence represents the minimum energy cost of physically running a simulation.

This result is complementary to the discussion of Crutchfield et al. who derive the difference between the two measures from the asymmetry of running the process in forward and reverse [7]. We add to this a physical interpretation of the cost of reversing a computation using thermodynamics. The lower bound to the energy cost of simulating a physical system was derived for optimal classical simulators. Recent results that quantum simulators require less information storage could indicate that quantum information leads to a reduced cost for stochastic computation [20]. Our results reveal an intricate relation between thermodynamics, information processing, and complexity. They motivate the use of information-theoretic tools for studying the physics of complex systems.

* Electronic address: k.wiesner@bristol.ac.uk

[1] Charles H. Bennett. *Int. J. Theo. Phys.*, 21(12):905–940, 1982.

[2] K. Maruyama, F. Nori, and V. Vedral. *Rev. Mod. Phys.*, 81(1):1–23, 2009.

[3] Oliver Penrose. *Foundations of Statistical Mechanics*. Elsevier, 1970.

[4] W. H. Zurek. Thermodynamic cost of computation, algorithmic complexity and the information metric. *Nature*, 341(6238):119–124, 1989.

[5] J. P. Crutchfield and K. Young. *Phys. Rev. Lett.*, 63(2):105, July 1989.

[6] W. Bialek, I. Nemenman, and N. Tishby. *Neural Comput.*, 13(11):2409–2463, 2001.

[7] J. P. Crutchfield, C. J. Ellison, and J. R. Mahoney. arxiv/0902.1209, 2009.

[8] C. R. Shalizi and J. P. Crutchfield. *J. Stat. Phys.*, 104(3):817–879, 2001.

[9] J. P. Crutchfield and D. P. Feldman. *Chaos*, 13(1):25–54, 2003.

[10] J. P. Crutchfield and D. P. Feldman. *Phys. Rev. E*, 55(2):R1239, 1997.

[11] D. P. Varn, G. S. Canright, and J. P. Crutchfield. *Phys. Rev. B*, 66(17):174110, 2002.

[12] C.-B. Li, H. Yang, and T. Komatsuzaki. *P. Natl. Acad. Sci. USA*, 105(2):536–541, 2008.

[13] A.J. Palmer, C.W. Fairall, and W.A. Brewer. *IEEE T. Geosci. Remote*, 38(4):2056–2063, 2000.

[14] C. R. Shalizi, K. L. Shalizi, and R. Haslinger. *Phys. Rev. Lett.*, 93(11):118701, 2004.

[15] R. Landauer. *IBM J. Res. Dev.*, 5:191, 183, 1961.

[16] Angluin Dana. On the complexity of minimum inference of regular sets. *Information and Control*, 39(3):337–350, December 1978.

[17] P. Gaspard. *J. Stat. Phys.*, 117:599–615, 2004.

[18] L. Szilard. *Z. Phys.*, 53:840–856, 1929.

[19] Cosma Rohilla Shalizi and Cristopher Moore. What is a macrostate? subjective observations and objective dynamics. *cond-mat/0303625*, 2003.

[20] M. Gu, K. Wiesner, E. Rieper, and V. Vedral. Sharpening occam's razor with quantum mechanics. *arxiv/1102.1994*, 2011.

[21] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. WileyBlackwell.

[22] For definitions of Shannon entropy, conditional entropy, and mutual information, see for example [21]